**scikit-learn reports  
資訊114 F7401254 張暐俊**

1. **改善決策樹分類模型**
   * **增加更多的輸入特徵**一張含有 文字, 螢幕擷取畫面, 字型 的圖片

     自動產生的描述

上圖為助教在hint中範例程式碼使用的三個特徵，而我在閱讀完train.cv以及在搜索每個欄位上方英文字幕代表意義時，找到網路上有人分析好的相關資料，我決定將我認為會提高準確率的特徵加入。一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

一開始我只有選擇Pclass，因為我先入為主的認為船艙的艙等對於存活率肯定是有影響的，但結果預測的結果打了我的臉，準確率並沒有太大的變化。但有趣的是，在我看完網路上他人分析的資料後，我再加入Embarked，也是就是登船的地點，配合著等等會說明的前處理，我的準確率從0.72來到了0.759。 此外，我也有嘗試加入SibSp跟Parch兩筆特徵，但反倒使我的準確率下降了，無論是兩者同時加入，抑或是個別加入SibSp或Parch。

* + **使用不同的前處理方法**

在加入不同特徵前，我先摸索的正是前處理方式，因為我在閱讀完train.csv檔案之後，我發現裡面的data有的有缺失，有的沒辦法直接做運算，有的需要處理完之後實作上才會更方便。我首先下手的是「age」的前處理。根據助教提供的四個strategy去做調整，結論是助教使用的median是最佳的選擇了，使用其他的話，準確率皆會降至0.72以下。那原因應該是因為乘客的年齡分佈並非那麼的集中，有許多的極端值，像是高齡老人或是嬰兒，也因此相較於mean來說，median或most\_frequency是來得較好，但在實測後，我依舊維持原樣，選擇了median。一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

接著我開始著手摸索類別型前處理，根據助教上課提到的方法有label encoder跟one-hot encoder，在研究了一下後，我決定使用one-hot encoder來做替換，因為我認為在性別上應該是平等的，可是使用label encoder的話，可能會造成產生距離的問題，也讓我推估是造成準確率無法提高的原因。在更改之後，有使得我的準確率從0.72提高至0.73，正式突破0.72大關。

接著我加入了Pclass跟embarked兩個欄位，因為Pclass已經是以數字顯示，所以我只有著手embarked的部分，這次我則選擇label encoder，因為根據前述，我在閱讀資料時發現，登船地點的不同的確是影響生存率的原因之一，所以我希望他們本身之間含有距離關係。

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

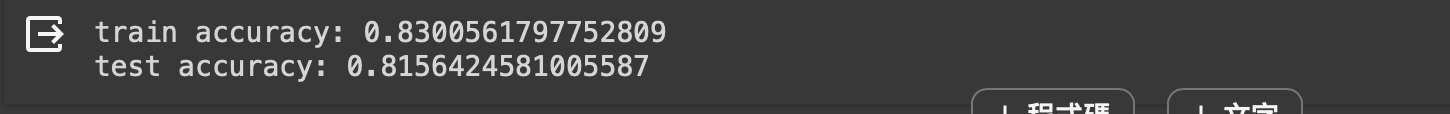
自動產生的描述

* + **調整超參數**一張含有 文字, 字型, 軟體, 螢幕擷取畫面 的圖片

    自動產生的描述

最後則是研究超參數，這邊我並沒有多做探索，我選擇增加的只有上圖的criterion、max\_depth跟max\_leaf\_node，相較於助教在課堂上使用的gini，我選擇使用entropy，相較於gini的隨機性來說，entropy的選擇更傾向於平均值，但他同樣有缺點，就是他的樹不能太深，否則其缺點同樣會很明顯。

因此在調整並測驗後，我選depth = 3當作我的樹深，我的test\_accuracy則是來到了新高的0.81。一張含有 文字, 螢幕擷取畫面, 字型, 軟體 的圖片

自動產生的描述 

* + **額外發現**

隨著調整上方三項，我的train\_accuracy也會隨之改變，尤其在我調整模型的超參數時最為明顯，也因此我有在想，儘管我使得我的test\_accuracy來到了0.81隻高，會不會我在選擇max\_depth = 4時，儘管test\_accuracy略低，但是train\_accuracy的數值反而較高，這樣子比較好呢？或許在兩者之間取得平衡，也是在調整這些參數或是方法時，需要去注意並思考的一環。一張含有 文字, 螢幕擷取畫面, 字型, 軟體 的圖片

自動產生的描述

1. **使用不同模型**
   * **測試結果**

|  |  |  |  |
| --- | --- | --- | --- |
| **模型類別** | **模型名稱** | **Train Accuracy** | **Test Accuracy** |
| sklearn.naive\_bayes | GaussianNB | 0.77528 | 0.80446 |
| BernoulliNB | 0.78370 | 0.79888 |
| sklearn.svm | LinearSVC | 0.77528 | 0.77653 |
| NuSVC | 0.79353 | 0.80446 |
| SVC | 0.78370 | 0.79888 |
| sklearn.ensemble | RandomForestClassifier | 0.91573 | 0.79329 |
| ExtraTreesClassifier | 0.83707 | 0.79329 |
| GradientBoostingClassifier | 0.90308 | 0.76536 |
| AdaBoostClassifier | 0.81179 | 0.81005 |
| sklearn.neural\_network | MLPClassifier | 0.77668 | 0.81005 |
| sklearn.linear\_model | SGDClassifier | 0.68398 | 0.74301 |
| LogisticRegression | 0.77808 | 0.80446 |

1. **參考資料**
   * 机器学习：泰坦尼克数据分析

**<https://zhuanlan.zhihu.com/p/447348987>**

* + Decision Trees Explained — Entropy, Information Gain, Gini Index, CCP Pruning

[**https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c**](https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c)

* + Different Types of Distance Metrics used in Machine Learning

**<https://medium.com/@kunal_gohrani/different-types-of-distance-metrics-used-in-machine-learning-e9928c5e26c7>**

* + Titanic Machine Learning by k-nearest neighbors (KNN) algorithm

[**https://medium.com/analytics-vidhya/titanic-machine-learning-by-k-nearest-neighbors-knn-algorithm-530d8bdd8323**](https://medium.com/analytics-vidhya/titanic-machine-learning-by-k-nearest-neighbors-knn-algorithm-530d8bdd8323)

* + 機器學習\_學習筆記系列(69)：K維樹(KD Tree)

[**https://tomohiroliu22.medium.com/機器學習-學習筆記系列-69-k維樹-kd-tree-b3b8591c9245**](https://tomohiroliu22.medium.com/機器學習-學習筆記系列-69-k維樹-kd-tree-b3b8591c9245)

* + 機器學習\_學習筆記系列(70)：球樹(Ball Tree)

[**https://tomohiroliu22.medium.com/機器學習-學習筆記系列-70-球樹-ball-tree-7bccb54cbf1e**](https://tomohiroliu22.medium.com/機器學習-學習筆記系列-70-球樹-ball-tree-7bccb54cbf1e)

* + [資料分析&機器學習] 第4.1講 : Kaggle競賽-鐵達尼號生存預測‑(前16%排名)

[**https://medium.com/jameslearningnote/資料分析-機器學習-第4-1講-kaggle競賽-鐵達尼號生存預測-前16-排名-a8842fea7077**](https://medium.com/jameslearningnote/資料分析-機器學習-第4-1講-kaggle競賽-鐵達尼號生存預測-前16-排名-a8842fea7077)

* + [資料分析&機器學習] 第3.4講：支援向量機(Support Vector Machine)介紹

[**https://medium.com/jameslearningnote/資料分析-機器學習-第3-4講-支援向量機-support-vector-machine-介紹-9c6c6925856b**](https://medium.com/jameslearningnote/資料分析-機器學習-第3-4講-支援向量機-support-vector-machine-介紹-9c6c6925856b)

* + scikit-learn中文社区

[**https://scikit-learn.org.cn**](https://scikit-learn.org.cn)

* + scikit-learn

[**https://scikit-learn.org/stable/index.html**](https://scikit-learn.org/stable/index.html)

* + 樸素貝葉斯（數值範例）

[**https://medium-com.translate.goog/@balajicena1995/naive-bayes-numerical-example-afcfa2433f95?\_x\_tr\_sl=en&\_x\_tr\_tl=zh-TW&\_x\_tr\_hl=zh-TW&\_x\_tr\_pto=sc&\_x\_tr\_hist=true**](https://medium-com.translate.goog/@balajicena1995/naive-bayes-numerical-example-afcfa2433f95?_x_tr_sl=en&_x_tr_tl=zh-TW&_x_tr_hl=zh-TW&_x_tr_pto=sc&_x_tr_hist=true)